The Heliophysics Data Policy: Status and Plans

Aaron Roberts
NASA GSFC
27 February 2012
Heliophysics Subcommittee Meeting

The Data Policy states various functions for the HP Data Environment

- Produce and serve high-quality, well-documented data
- Provide open access to scientifically useful data products
  - Allow easy discovery of all available products and their location
  - Provide easily useable, well-documented products
  - Provide uniformity of access to data
- Keep data flowing without interruption when missions end
  - Provide funds to continue post-mission serving of data, if needed
  - Move data to Active Final Archives for long-term serving
- Keep data safe for the long term
  - Assure data are safe at all stages
  - Provide long-term archives for safe-keeping

The HP Data Policy is working

- New missions are following PDMP guidelines and will deliver data as expected; VxOs and Final Archives are involved in the process.
- Current missions are improving their data, documentation, and services; most are in good shape.
- Senior Reviews and Mission Archive Plans continue to help.
- Data are moving into Active Final Archives, and are being served and kept safe.
- An Inventory and Registry of all HP data is being completed and has an active interface (VSPO) that will deliver or point directly to data.
- Legacy datasets are being improved, archived, and served.
- Plans are moving forward for uniform access to HP data.
  - HDMC/VxOs

Inventory/Registry: SPASE is stable and working

- "SPASE" is the "Data Model" that allows us to provide uniform descriptions ("metadata") for all Heliophysics data products and services

- Most data products from nearly 100 space-based and many more ground based observatories are registered using SPASE (includes 30 solar observatories, space- and ground-based)

- Nearly all available data from all NASA HP active missions is directly accessible
  - Easily discovered by time range, cadence, general region, measurement type, name, relevant text in description, person name, ... or any combination of the above.
  - Parameter range, magnetospheric state, spacecraft/ground coincidence, and/or event lists available for searching for some data products, depending on VxO.

- Non-NASA data largely accounted for (some availability and access problems)

- Lag in SPASE descriptions at the detailed (parameter/variable) level
  - Affects universality of access and limits some types of search
  - Being addressed

Problem of Uniform Access (asking for all data in the same way)

- Advantages of self-documenting, standard formats
  - Variable names, units, etc., are encoded in a uniform way
  - Time is in a fixed format and is thus instantly readable
  - Descriptive metadata is tied to the relevant variables
  - Internet access from, e.g., IDL or MatLab can be easily automated
  - *CDF-A (CDF + time, structuring, and metadata conventions) being developed to have a truly archival CDF*
  - *FITS and NetCDF (probably TIMED conventions) should complete our set*

- SPASE-based access (e.g., access by: SPASE ID; time range; variable 'keys')
  - Metadata required, but can be difficult to get (progress being made)
  - SPASE-QL and/or general Data Access Protocols use the metadata
  - Currently implemented by some VxOs and CDAWeb.
  - VxOs—possibly ultimately not so much portals as formulators and implementers of standards (VAO/IVOA path) for the protocols; general tools build on that.

- ASCII "problem"
  - Makes simple uses easy, but lack of standards means more metadata required for easy, direct access
  - Schemes for generating and using such metadata are being generated
  - Copies in standard formats can and do solve the problem

CDAWeb data directly to IDL

- To ***directly fill arrays from web sources using routines within IDL***:
- Download the file “spdfcdas.sav” [contains all needed CDF routines]
- In IDL:
  - restore, “spdfcdas.sav”
  - uly1sec95_6 = spdfgetdata('UY_1SEC_VHM', ['B_RTN', 'B_MAG'], ['1995-06-29T00:00:00.000Z', '1995-06-31T00:00:00.000Z'])
  - structid = ‘uly1sec95_6’ [the name of the ‘structure’ with everything in it]
  - assign_variables [Pulls the variables out of the structure and gives them names according to the CDF metadata]
  - [Can also invoke the following for a gui dataset/variable/time range chooser:]
  - spdfcdawebchooser [allows direct reading or command generation]
  - [IDL reports back the names of the variables that have been read in.]
  - qq = plotmaster(uly1sec95_6,/auto) [will plot all the data as in CDAWeb]
  - [A general routine exists to put variables on a uniform time basis by averaging or interpolating as needed.]
- VxOs are working on generalizing this and similar capabilities to distributed data from multiple sources.

Most datasets are now safe for the long-term and actively served

- Science-quality, high-resolution data are at SPDF (CDAWeb or ftp), in most cases for most or all instruments:
  - ACE, Wind, Polar, IBEX, Voyager, Pioneer, Helios, THEMIS, STEREO (in situ), Ulysses, SOHO (particles), Geotail, IMP-8, DE, Many Explorers, ISIS, TWINS, Cluster (prime parameters), some others being negotiated. Also OMNI.
  - Other countries also preserve data (notably, the Cluster Active Archive, but also many others such as Akebono and Geotail at DARTS).
  - RAs keep IMAGE/RPI, FAST, many Polar datasets, and other space physics data flowing.
  - Long-term backups via NSSDC

- There are many active solar missions (Hinode, SOHO, SDO, RHESSI, STEREO imaging); data are well served and probably quite safe, e.g., with copies served from SDAC, but not as clear a plan in some cases. RAs exist for a number of older missions (TRACE, Yohkoh, SOHO MDI), and other countries also preserve data (e.g., Hinode at DARTS; SOHO and RHESSI in Europe).

- Probably safe, but no NASA plan: IMAGE ENAs, SAMPEX, non-NASA (DMSP, NOAA, etc.)

Datasets being restored/improved/upgraded

- ISEE-1, -2, FAST, WIND/SWE, SUSIM irradiance, Mees Vector Magnetograms, DE-1 plasma waves, SMM Gamma-rays, RHESSSI standardized images, etc.

- We are reaching the end of the list of useful cases
  - New proposals tend to be for more subtle improvements rather than basic restoration.
  - Remaining known datasets (e.g., at NSSDC) currently in nonstandard form are typically older, shorter, and “less interesting.”
  - There may be some things we just cannot afford although they would be useful, but not many.

Future challenges/vision

- Metadata production and use
  - Definitive inventory/registry: referential (DOI?) and discovery uses
  - Uniform data access for all products (VxO focus)
  - Seamless flow from mission archives through to final archives
- Format standards (e.g., CDF-A; also NetCDF standard?)
  - Adoption of standards in calls for mission proposals (the time has come!)
- Large data volumes
  - How to use the data: Pattern recognition; data mining
  - How to keep the data available and safe post-mission
- Model-data comparisons and insights
  - Seamless integration of model output with data streams
  - Data assimilation; true space weather capabilities
  - Data volume questions, as above
- VxO plans
  - Complete the Inventory/Registry of data products
  - Complete discovery and access tools and basic visualization based on them
  - Consolidate to core groups in Solar, ITM, and Helio/Magnetosphere; determine level of resources needed
  - Focus on providing metadata and links to generic access methods, especially for new missions